DOCUMENT RESUME

ED 194 579 TM 800 708

AUTHOR Grandy, Jerilee
TITLE Analysis of the Subscale Structure of Test Batteries: A Confirmatory Study of the Interrelationships of CGP and N.J. Basic Skills Subscores.
INSTITUTION Educational Testing Service, Princeton, N.J.
REPORT NO ETS-RR-80-25
PUB DATE Oct 80
NOTE 24p.

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS College Bound Students; *College Entrance Examinations; College Freshmen; Correlation; Criterion Referenced Tests; *Factor Structure; *Goodness of Fit; Higher Education; Placement; *Test Validity
IDENTIFIERS *Comparative Guidance and Placement Program; Confirmatory Factor Analysis; *New Jersey College Basic Skills Placement Test

ABSTRACT

The primary purposes of the study were (1) to investigate whether three subtests of the Comparative Guidance and Placement Program (CGP): Mosaic Comparisons, Year 2000 and Letter Groups measure skills uniquely different from traditional verbal and mathematical skills; and (2) to test whether the New Jersey College Basic Skills Placement Test subtests are measuring the same skills as similarly named subtests of the CGP. The methodology employed was a confirmatory factor analysis using the COFAMM computer program. Data from 822 students who had taken both batteries were used to test a hypothesized four-factor model (Reading, Sentences, Mathematics, and Mosaic Comparisons). This model was found to fit the data; that is, subtests with the same names measured the same skills. Basic Skills subtests, Mosaic Comparisons, Year 2000, and Letter Groups, however, each measured something uniquely different from Reading, Sentences, and Mathematics. Although technically complex, this methodology is easily and inexpensively applied to this type of problem. It can be particularly useful in criterion-referenced test development for testing whether a priori subscales are actually measuring different skills. (Author/CP)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED194579

RR-80-25

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

RESEARCH REPORT

ANALYSIS OF THE SUBSCALE STRUCTURE OF TEST BATTERIES: A CONFIRMATORY STUDY OF THE INTERRELATIONSHIPS OF CGP AND N.J. BASIC SKILLS SUBSCORES

Jerilee Grandy

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
Educational Testing Service
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

TM 800 708

Educational Testing Service
Princeton, New Jersey
October 1980

ERIC

ANALYSIS OF THE SUBSCALE STRUCTURE OF TEST BATTERIES: A CONFIRMATORY STUDY OF THE INTERRELATIONSHIPS OF CGP AND N.J. BASIC SKILLS SUBSCORES

Jerilee Grandy

Educational Testing Service
Princeton, New Jersey
October 1980

ERIC

Copyright © 1980. Educational Testing Service. All rights reserved.

Abstract

This analysis exemplifies a method for investigating the construct validity of the subscales of one or more test batteries. In this study, the structures of the Comparative Guidance and Placement Test (CGP) and New Jersey Basic Skills test were examined to see if the subscales designed to measure the same skills were doing so, and to see if the nontraditional subscales (Mosaic Comparisons, Year 2000, and Letter Groups) were measuring something different from the traditional subscales (Reading, Sentences, and Mathematics).

The methodology employed was a confirmatory factor analysis. Data from 822 students who had taken both batteries were used to test a hypothesized four-factor model (Reading, Sentences, Mathematics, and Mosaic Comparisons); this model was found to fit the data. It was concluded that Mosaic Comparisons, Year 2000, and Letter Groups each measure something uniquely different from Reading, Sentences, and Mathematics.

Although technically complex, this methodology is easily and inexpensively applied to this type of problem. It can be particularly useful in criterion-referenced test development for testing whether a priori subscales are actually measuring different skills.

Introduction

The Comparative Guidance and Placement (CGP) test battery is designed to measure skills in reading, sentence structure, and mathematics plus the skills required to do three subtests entitled Mosaic Comparisons, Letter Groups, and Year 2000. A question had arisen as to whether the three non-traditional subtests were actually measuring something different from the standard verbal and mathematics tests. A further question was also raised in connection with the New Jersey Basic Skills test. This test, also produced by ETS, has subtests for reading, sentences, and mathematics. Do these subtests measure the same skills as the CGP subtests bearing similar names?

These questions deal with complementary aspects of construct validity, namely, convergent and discriminant validity (see Cronbach, 1971, and Campbell and Fiske, 1959). A test has convergent validity if it measures what it purports to measure. It has discriminant validity if the skill it measures is distinctly different from other skills.

Recent developments in maximum likelihood confirmatory factor analysis enable the researcher to test, statistically, the goodness-of-fit of *a priori* construct validation models to empirical data. (See Werts & Linn, 1970, for the application of path analysis techniques to the multitrait-multimethod matrix; see Rock & Werts, 1979, for a recent example.) When the fit of a given model is not rejected, it can be concluded that some evidence of the construct validity of the measures has been found and that the underlying theory of the interrelationships of the variables has been, to some extent, corroborated. When the relationships involving true scores are modeled with confirmatory factor structures, a number of psychometric parameters can be estimated. As Jöreskog (1971) points out, for the scores that are congeneric (i.e., measures of the same factor), the maximum likelihood estimates of the factor loadings are the regression of the observed scores on their "true" scores. Squared standardized factor loadings correspond to estimates of the reliability with which each instrument measures each skill (construct or factor). The correlation between factors corresponds to the correlation between variables corrected for attenuation, i.e., the correlation between true scores (see Werts & Linn, 1972).

Data

When students take the CGP, they complete four separately timed reading sections, four sections of sentences, three sections of Mosaic Comparisons, one section entitled Letter Groups, and one section called Year 2000. For the mathematics tests, they are instructed to take math level C if they have had no algebra in high school. Level C consists of two sections; one is computation and the other consists of arithmetic reasoning. Students who have had one year of algebra in high school are told to take level D. This consists of the same computation test included in level C and an elementary algebra test. Level E, which is taken by those who have had two years of high school algebra, consists of the same elementary algebra test plus an intermediate algebra test. Thus all students take two mathematics tests.

In the New Jersey Basic Skills Test, there are three reading subtests, three sentence subtests, one computation test, and one elementary algebra test. All of these are taken by all students.

Data from 822 students who had taken both the CGP and the New Jersey Basic Skills Tests were used for analysis. Scores were divided into three

ERIC

groups corresponding to the level of CGP Mathematics taken. Level C consisted of 184 students, level D had 282 students, and level E consisted of 356 students.

Method

A model was developed in which Reading, Sentences, Mosaic Comparisons, Math Computation, and Elementary Algebra were hypothesized as five constructs (factors) underlying the scores being studied. The four Reading scores on the CGP and three Reading subscores of the New Jersey test were hypothesized to load on a single "Reading" factor. This factor would be interpreted as "true" reading skill. Likewise, the four CGP sentences subscores and the three New Jersey Sentences subscores were hypothesized to fit a single factor labeled "Sentences."

Because it was not certain whether Math Computation and Elementary Algebra would fit a single math factor, and because there was no *a priori* necessity that they do so, two different math factors were hypothesized. One measure of each factor, Computation and Elementary Algebra, arose from each test battery.

The three sections of Mosaic Comparisons were hypothesized to form a correlated but distinctly different factor from the other four.

Since there was only one score for Year 2000 and one for Letter Groups, these measures could not be treated as separate factors. Instead, they were permitted to load on all other factors. By permitting them to do so, estimates could then be made of the degree to which scores on those subtests could be explained by each of the factors (i.e., each of the traditional subject areas).

The COFAMM program (Sörbom & Jöreskog, 1976) was used to test the fit of this model to the data, simultaneously for each of the three subgroups of students, as defined by the three math subtest levels. COFAMM assumes that a factor analysis model holds in each of the g populations under study. If $\underset{\sim}{x}_g$ is defined as the vector of the p observed measures in group g, then $\underset{\sim}{x}_g$ can be accounted for by k common factors ($\underset{\sim}{f}_g$) and p unique factors ($\underset{\sim}{z}_g$). The model in each population is:

$$\underset{\sim}{x}_g = \underset{\sim}{\nu}_g + \underset{\sim}{\Lambda}_g \underset{\sim}{f}_g + \underset{\sim}{z}_g \qquad (1)$$

where $\underset{\sim}{\nu}_g$ is a p x 1 vector of location parameters and $\underset{\sim}{\Lambda}_g$ a p x k matrix of factor loadings. It is assumed that $\underset{\sim}{z}_g$ and $\underset{\sim}{f}_g$ are uncorrelated, the expectation of $\underset{\sim}{z}_g = 0$ and the expectation of $\underset{\sim}{f}_g = \underset{\sim}{\theta}_g$ where $\underset{\sim}{\theta}_g$ is a k x 1 parameter vector.

Given these assumptions, the mean vector $\underset{\sim}{\mu}_g$ of the $\underset{\sim}{x}_g$ is

$$\underset{\sim}{\mu}_g = \underset{\sim}{\nu}_g + \underset{\sim}{\Lambda}_g \underset{\sim}{\theta}_g \qquad (2)$$

and the expected variance-covariance matrix $\underset{\sim}{\Sigma}_g$ of $\underset{\sim}{x}_g$ is

$$\underset{\sim}{\Sigma}_g = \underset{\sim}{\Lambda}_g \underset{\sim}{\phi}_g \underset{\sim}{\Lambda}_g' + \underset{\sim}{\Psi}_g \qquad (3)$$

where $\underset{\sim}{\phi}_g$ is the variance-covariance matrix of the $\underset{\sim}{f}_g$ and $\underset{\sim}{\Psi}_g$ is the variance-covariance matrix of $\underset{\sim}{z}_g$. When the factor model does not fit the data perfectly, the observed variance-covariance matrices $\underset{\sim}{S}_g$ and observed means will differ from the maximum likelihood estimates of $\underset{\sim}{\Sigma}_g$ and $\underset{\sim}{\mu}_g$. The program yields a chi-square statistic that is a measure of these differences, that is, of how well the hypothesized model fits the data compared to the null hypothesis that the variance-covariance matrix of $\underset{\sim}{z}_g$ may have any structure whatsoever.

The four matrices Θ_g, Λ_g, Φ_g, and Ψ_g are called the pattern matrices. The elements of these matrices are the model parameters which are of three kinds: (a) fixed parameters, which have been assigned given values, like 0 or 1; (b) constrained parameters, which are unknown but equal to one or more other parameters; and (c) free parameters, which are unknown and not constrained to be equal to any other parameter. A parameter may be constrained to be equal to other parameters in the same and/or different pattern matrices in the same and/or in different groups.

The important feature of a confirmatory analysis is that the parameters of the model may be uniquely estimated, i.e., the model is identified subject to an algebraic constraint. According to this constraint, a solution is unique if all linear transformations of the factors that leave the fixed parameters unchanged also leave the free parameters unchanged. It is difficult in general to give useful conditions which are sufficient for identification. However, at one point in the program the information matrix for the unknown parameters is computed. If this matrix is positive definite, it is almost certain that the model is identified. If this matrix is not positive definite, the program prints a message to this effect, specifying which parameter is probably not identified.

In this study, the model is overidentified, yielding not only unique solutions but sufficient degrees of freedom for statistical tests of goodness-of-fit. In addition, standard errors for all the unknown parameter estimates are also provided by the program. This analysis differs from an exploratory factor analysis. In an exploratory analysis, the model is usually not identified, and thus, there is neither a statistical test of goodness-of-fit nor is there a unique solution.

Results and Discussion

The factor pattern of the initial model to be tested was a special case of equation (1), with 23 variables and 5 hypothesized factors. These factors were Reading, Sentences, Computation, Algebra, and Mosaic Comparisons. The structure being tested was hypothesized to be the same across all three groups of subjects. The statistical test of this model using COFAMM yielded a chi-square of 994.56 with 720 degrees of freedom. Considerable difficulty was encountered in finding a solution for this model, however, because of a high colinearity between the two math factors ($r = .995$). This is because regression weights on highly correlated independent variables have large standard errors (see Farrar & Glauber, 1967).

The model was, therefore, reduced to four factors with Computation and Elementary Algebra scores being permitted to load on a single "Mathematics" factor. When tested, this model yielded a chi-square of 1020.5 with 721 degrees of freedom and an RMS of the residuals of .07.

Because of the large sample size, even the most trivial deviations from the model would tend to yield a statistically significant chi-square value. This chi-square, however, is relatively small compared to the degrees of freedom. A more appropriate measure of goodness-of-fit is the root mean square (RMS) of the residuals. This is the square root of the average squared difference between corresponding elements in each population's observed variance-covariance matrix and the reproduced variance-covariance matrix conditional on the constrained factor model. Ordinarily, it would not be easy to interpret these indices in the case of variance-covariance matrices, therefore, the original variance-covariance matrices were rescaled to correlation matrices (e.g., see Sörbom & Jöreskog, 1976). Thus, the RMS may be interpreted much as you would interpret the residuals when fitting a factor model to the observed correlation matrix.

The RMS of .07 found here is satisfactory considering the sample sizes and differences between the three mathematics ability groups. Table 1 shows the standardized factor loadings and percentage of variance in each observed score that can be explained by each factor. Table 2 shows the intercorrelation of the four factors. Standard errors for these estimates were about .03 for all factor loadings with the exception of those for Year 2000 and Letter Groups which had standard errors of .06 for their loadings on the Reading and Sentences factors. Standard errors were .04 in the correlations between factors, except for the correlation between Reading and Sentences which had a standard error of only .02.

A diagram of the model is shown in Figure 1 with standardized factor loadings indicated. To simplify the drawing, intercorrelations of factors are shown separately in Figure 2. The circles represent factors (constructs or true scores), and arrows from the factors are shown to indicate that those true scores underlie the observed scores to which they point. The small arrows indicate the measurement error component of the observed score. These figures merely show diagrammatically the same information contained in the tables.

The factor intercorrelations can be interpreted as correlations between true scores--i.e., correlations corrected for attenuation. Mosaic Comparisons, when corrected for attenuation, correlate only slightly with Reading and Sentence skills (.20 and .26) but somewhat more

ERIC

moderately with Mathematics (.37). These results suggest that although there is some relationship between the abilities required to do Mosaic Comparisons and the abilities required to do traditional verbal and math tests, there are other unique skills required to do Mosaic Comparisons.

The standardized factor loadings obtained for each subtest score are most useful when squared and interpreted as percentages of variance explained by the underlying factor. Because Year 2000 and Letter Groups loaded on all four intercorrelated factors, their factor loadings are partial correlations. It can be seen that very little variance in either score is explained by Reading or the skill underlying Mosaic Comparisons. For Year 2000, 12% of the variance is explained by the Sentences factor and 8% by Math skill. The combination of all four factors, however, accounts for 54% of the variance in Year 2000. (This number is obtained by subtracting the unknown unique variance from unity rather than by adding the squared standardized factor loadings because, for Year 2000 and Letter Groups these are partial factor loadings.)

For Letter Groups, 10% of the variance can be explained by Math skill and 23% by Sentences. Note that the Letter Groups score is as reliable a measure of the sentences factor as is the first CGP Sentences subscore. Nevertheless, a considerable amount of variance in the Letter Groups scores remains unexplained, the four factors explaining only 40% of the variance.

Pictorially, the effects of the four factors on Year 2000 and Letter Groups scores can be shown more clearly than in Figure 1 by recourse to two other figures excerpted from Figure 1. Figures 3a and 3b show the observed scores as functions of four factors plus uniqueness (Ψ).

It may be concluded from this study that while Mosaic Comparisons, Year 2000, and Letter Groups each have some variance explainable by the traditional verbal and mathematical test scores, each has a considerable proportion of unique variance (Ψ), uncorrelated with reading and math, that cannot be identified on the basis of these data. The question of whether these unique skills are related to college performance remains to be answered.

In addition to these conclusions, a number of other interpretations based on this model are notable. Because the model fit the data reasonably well, this analysis provides some evidence of construct validity for the Reading, Sentences, and Math tests in both test batteries. The fact that all subscales loaded on the expected factors supports the hypothesis that they are measuring what they were designed to measure. The Reading subscales, for example, all measure Reading rather than, say, Sentences. We might have discovered, for example, that CGP Reading subscale 4 loaded on the Sentences factor rather than the Reading factor. Or, we might have found that the New Jersey tests had a different Reading factor than the CGP did. The fact that this did not occur lends support to the validity of the Reading subtests. Likewise, the math subtests from both batteries loaded on a Math factor rather than on a Reading factor or on two different Math factors. The fact that this model fit as hypothesized provides evidence of both convergent and discriminant validity (and hence, construct validity) for the traditional subtests of both batteries.

Another point of interest is that in the analysis of those students who took CGP Math test C (i.e., those who had no Algebra course), the factor loading for the New Jersey Algebra scores was found to be only .16. This finding is consistent with the presupposition that an Algebra test cannot adequately measure knowledge of Algebra in a group of students who never studied it. For many students it may be measuring how well they figure out a problem by some other method or how clever they are. The factor loadings for Algebra reported in this paper are based only on data from students taking CGP Math test D or E.

Also of interest is the finding that the Computation and Algebra subtests from both batteries loaded on a single Math factor. The fact that this occurred supports the hypothesis that Algebra is just a harder form of Mathematics and is not a different construct. The finding provides support for any attempt that might be made to equate the two.

Similarly, it is possible on the basis of this analysis to calibrate all subtests loading on the same factor (Werts, Grandy, & Schabacker, 1980). The analysis supports the hypothesis that the CGP Reading subtests, for example, are measuring the same construct as the Reading subtests of the New Jersey Test. It is justifiable, therefore, to calibrate the New Jersey battery to the CGP battery.

Summary of Major Findings

The primary purposes of the study were twofold: (1) to investigate whether certain subtests of the CPG (Mosaic Comparisons, Year 2000, and Letter Groups) are measuring skills uniquely different from traditional verbal and mathematical skills; (2) to test whether the New Jersey Basic Skills subtests are measuring the same skills as similarly named subtests of the CGP. A five-factor model (Reading, Sentences, Mosaic Comparisons, Computation, and Algebra) was hypothesized and found to fit the data from the two test batteries. Those tests having the same names were found to be measuring the same skills. It was concluded, therefore, that both batteries have convergent validity. Two of the factors -- Math

Computation and Algebra were so highly correlated that they were inseparable. Estimates were, therefore, obtained from a model consisting of four factors -- Reading, Sentences, Mosaic Comparisons, and Mathematics. The skill required to do Mosaic Comparisons was found to correlate moderately ($r = .37$) with Mathematics and less well with the other two factors. Scores on Year 2000 loaded most heavily on the Sentences factor, as did the scores on Letter Groups. Letter Groups was, in fact, found to be as reliable a measure of the factor underlying Sentences as the first of the Sentences subscores. On the other hand, only 40% of the variance in Letter Groups could be explained by the four factors. The variance in Year 2000 scores was found to be 54% explained by the four factors.

It was concluded that while verbal and mathematical skills can, to some extent, account for a student's performance on Mosaic Comparisons, Year 2000, and Letter Groups, each is also measuring something distinctly different. They may, therefore, be said to have discriminant validity, and hence, construct validity. Whether the unique skills underlying these measures are relevant to college performance could not be ascertained from the existing data.

Table 1

Standardized Factor Loadings and Percentage of
Variance (in Parentheses) by Each Factor

Subscore	Factor: Reading	Factor: Sentences	Factor: Mathematics	Factor: Mosaic Comparisons
CGP Reading I (Main Idea)	0.62 (38%)			
CGP Reading II (Secondary Idea)	0.79 (62%)			
CGP Reading III (Inferences)	0.76 (58%)			
CGP Reading IV (Vocabulary)	0.81 (66%)			
NJ Reading I (Main Idea)	0.76 (58%)			
NJ Reading II (Direct Statements)	0.72 (52%)			
NJ Reading III (Inferences)	0.75 (56%)			
CGP Sentences I (Idiom and Diction)		0.50 (25%)		
CGP Sentences II (Coordination and Subordination)		0.63 (40%)		
CGP Sentences III (Agreement and Reference)		0.67 (45%)		
CGP Sentences IV (other)		0.74 (55%)		
NJ Sentence Structure I (Complete Sentences)		0.75 (56%)		
NJ Sentence Structure II (Coordination and Subordination)		0.76 (58%)		
NJ Sentence Structure III (Placing Modifiers)		0.74 (55%)		

ERIC

Table 1 (cont'd.)

Standardized Factor Loadings and Percentage of Variance (in Parentheses) by Each Factor

Subscore	Factor: Reading	Sentences	Mathematics	Mosaic Comparisons
CGP Math Computation			0.83 (69%)	
CGP Elementary Algebra			0.84 (71%)	
NJ Math Computation			0.75 (56%)	
NJ Elementary Algebra			0.72 (52%)	
CGP Mosaic Comparisons I				0.78 (61%)
CGP Mosaic Comparisons II				0.90 (81%)
CGP Mosaic Comparisons III				0.78 (61%)
CGP Year 2000*	0.15 (2%)	0.35 (12%)	0.14 (2%)	0.29 (8%)
CGP Letter Groups**	-0.21 (4%)	0.48 (23%)	0.22 (5%)	0.32 (10%)

*Total variance explained by all four factors = 54%.
**Total variance explained by all four factors = 40%.

ERIC

Table 2

Intercorrelations of the Four CGP - New Jersey Basic Skills Factors

	Reading	Sentences	Mathematics	Mosaic Comparisons
Reading	1.00			
Sentences	0.83	1.00		
Mathematics	0.50	0.48	1.00	
Mosaic Comparisons	0.20	0.26	0.37	1.00

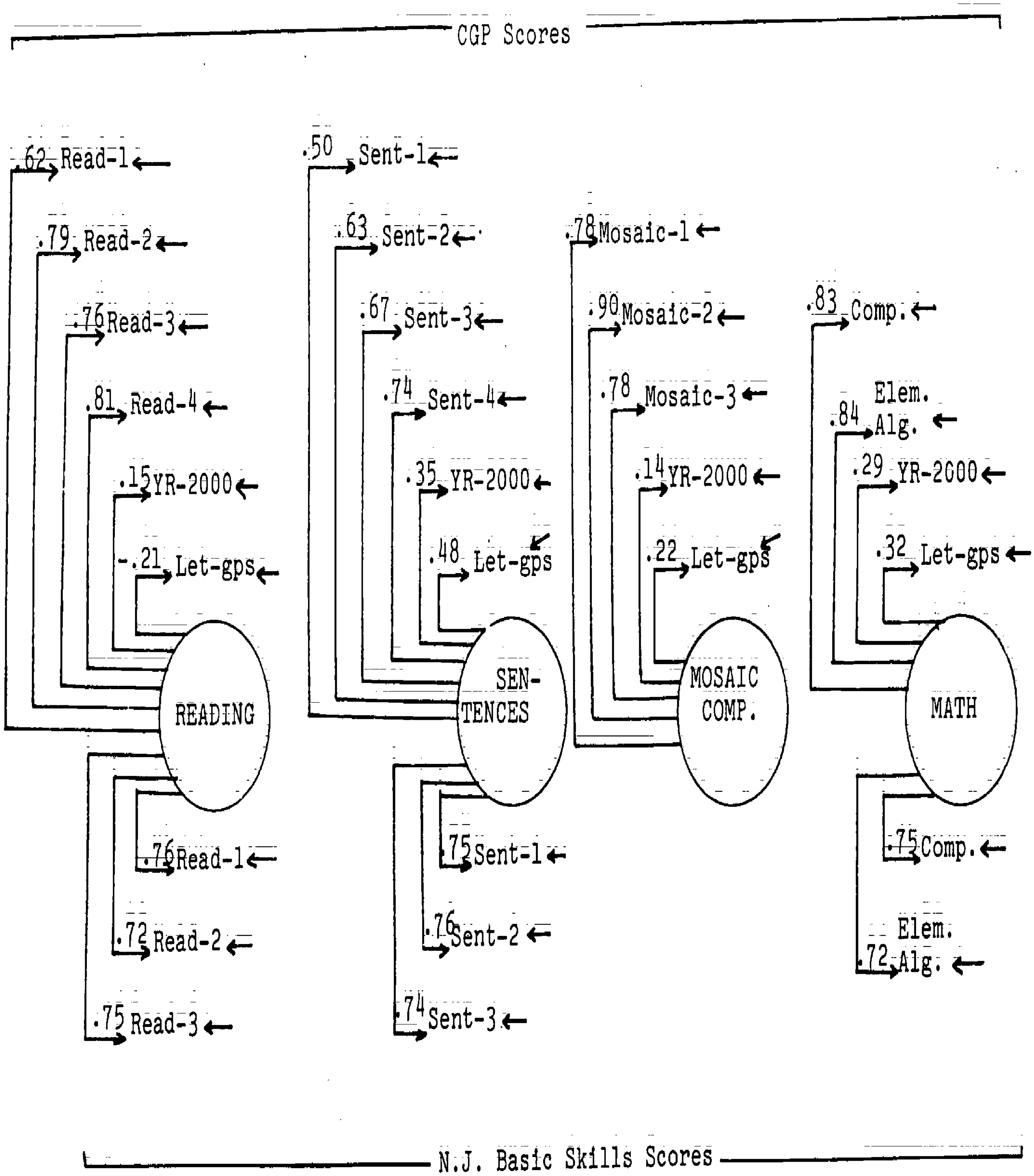

Figure 1. Confirmatory factor analysis model of CGP and N.J. Basic Skills Test scores. (Intercorrelations of factors not shown.)

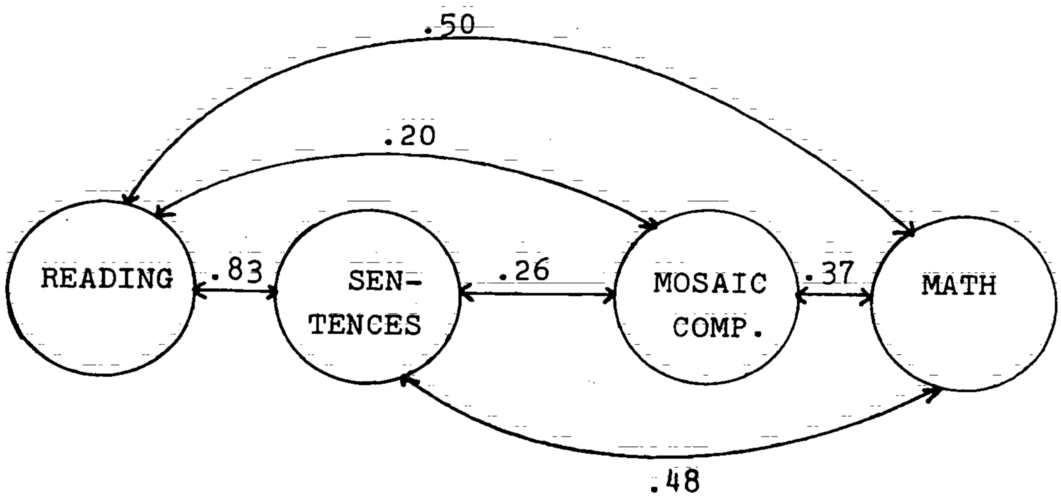

Figure 2. Intercorrelation of factors in model of CGP and N.J. Basic Skills Test scores.

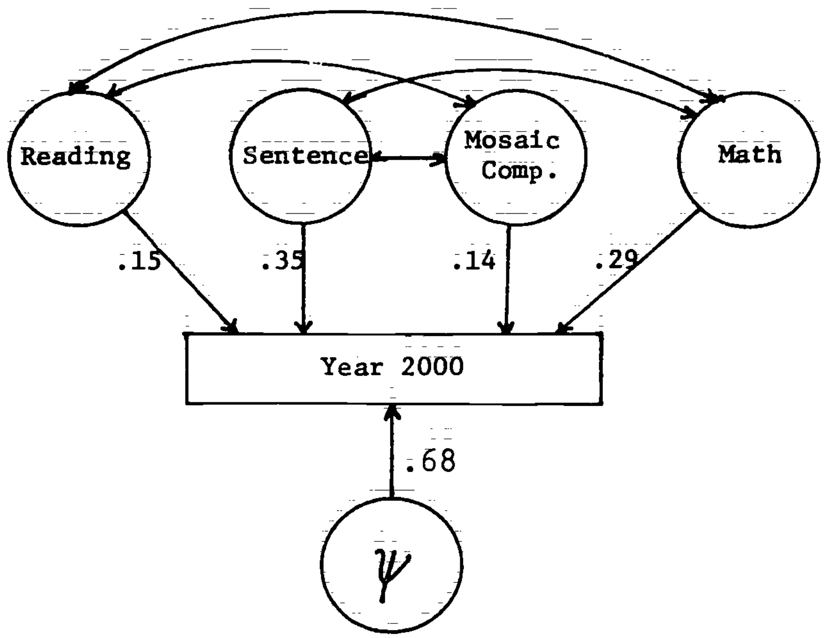

Figure 3a. Standardized partial factor loadings for Year 2000 scores.

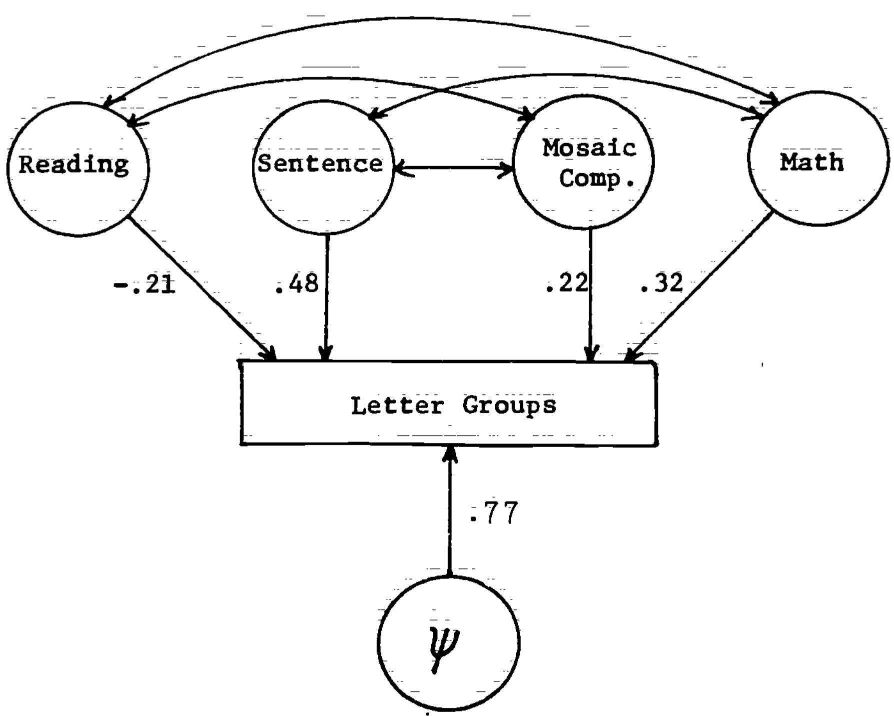

Figure 3b. Standardized partial factor loadings for Letter Groups scores.

References

Campbell, D., & Fiske, D. Convergent and discriminant validation by the multitrait-multimethod matrix. Psychological Bulletin, 1959, 56, 81-105.

Cronbach, L. J. Test validation. In R. Thorndike (Ed.), Educational Measurement. Washington, D.C.: American Council on Education, 1971.

Farrar, D. E., & Glauber, R. R. Multicollinearity in regression analysis: the problem revisited. Review of Economics and Statistics, 1967, 49, 92-107.

Jöreskog, K. G. Statistical analysis of sets of congeneric tests. Psychometrika, 1971, 36, 109-133.

Rock, D. A., & Werts, C. E. Construct validity of the SAT across populations--an empirical confirmatory study (RR 79-2). Princeton, N. J.: Educational Testing Service, 1979.

Sörbom, D., & Jöreskog, K. G. COFAMM: Confirmatory factor analysis with model modification (User's Guide). Chicago: National Educational Resources, Inc. 1976.

Werts, C. E., Grandy, J., & Schabacker, W. H. A Confirmatory factor approach to calibrating congeneric measures. Multivariate Behavioral Research, 1980, 15, 109-122.

Werts, C. E., & Linn, R. L. Path analysis: Psychological examples. Psychological Bulletin, 1970, 74, 193-212.

Werts, C. E., & Linn, R. L. Corrections for attenuation. Educational and Psychological Measurement, 1972, 32, 117-127.